APPLICATION OF K-MEANS CLUSTERING IN GROUPING CITIZEN WELFARE BASED ON SUB-DISTRICTS IN SURAKARTA CITY

Anggit Daneswara Purbaningrum [1], Diana Rahmawati [2], Diki Aryo Wijanarko [3], Khairina Altaf Salsabila [4], and Tiara Permata Sari[5]

[1,2,3,4,5]Department of Mathematics, Faculty of Mathematics and Natural Sciences,
Universitas Sebelas Maret, Indonesia,
[1]anggitdeswara@student.uns.ac.id,
[2]dianarahmawati01@student.uns.ac.id,
[3]dikiwijanarko@student.uns.ac.id,
[4]khairinaalsa@student.uns.ac.id,
[5]tiarapermata2806@student.uns.ac.id.

Abstract

The purpose of this article is to classify the welfare of sub-districts in Surakarta City using the *K-means Clustering* method. *The K-means Clustering* method has the ability to group large enough data in a very fast and efficient time. This method is included in the partition method which is based on the center point. This algorithm requires three parameters, namely the number of *clusters*, *cluster* initialization, and system distance. The data is taken from Surakarta City in Figures 2021, and the data loaded is 2020 data.

Introduction

1.1 Background of the Problem

Surakarta is one of the cities in Central Java. The palace, batik, and Klewer Market are three things that symbolize the identity of Surakarta. The existence of Surakarta Hadiningrat Kasunanan Palace and Mangkunegaran Temple (since 1745) makes Solo an axis, of history, art, and culture. Surakarta City has a population of 522,364 people spread across 5 sub-districts. The population growth rate is 0.44 percent. There are many inequalities in the masing kecamatan welfare of the people of each sub-district. Each sub-district has its own advantages, in terms of population, education, agricultural products, trade products, etc. To categorize the sub-districts in Surakarta City based on several aspects and factors. The welfare of the community in an area is a measure of progress in the region. We cluster with *K-means Clustering* method. The data is taken from Surakarta City in Figures 2021, and the data contained is 2020 data.

1.2 Problem formulation

Based on the background that has been described, the problem formulation is obtained, namely how to group the welfare of sub-districts in Surakarta City based on 2020 data using the *K-means Clustering* method.

1.3 Objectives

The purpose of this article is to group the welfare of sub-districts in Surakarta City using the *K-means Clustering* method.

1 Basic Theory

2.1 *Clustering* Methods

One of the techniques or methods in grouping data is *clustering*. *Clustering* is included in *unsupervised* data mining. One *cluster* usually has similar characteristics, and different *clusters* usually have very little or no similarity. So, the utilization of this *clustering* is to group data in a group.

A high degree of similarity within a *cluster* can be referred to as the result of a good *clustering* process. The data types in *clustering* are:

1. interval-scale variable,
2. binary variable,
3. nominal, ordinal, and ratio variables, and
4. variables with other types.

Clustering requirements that must be met in order to run the *clustering* algorithm are as follows:

1. scalability,
2. ability to analyze various forms of data,
3. determine *clusters* with unexpected shapes,
4. ability to be able to handle noise,
5. sensitivity to changes in inputs,
6. capable of *clustering* high-dimensional data, and
7. interpretation and usability.

Clustering methods have two types, namely hierarchical or hierarchical, and *non-hierarchical* or non-hierarchical [2]. In the hierarchical method, the number of groups to be formed has not been determined. Whereas in the non-hierarchical method, the *cluster* that is formed is determined in advance, so that objects will be grouped into k predetermined groups. Methods that are often used are *K-Means* and *Fuzzy C-Means*.

2.2 *Clustering* Methods

The K-means Clustering method has the ability to cluster large amounts of data very quickly and efficiently. This method belongs to the partitioning method which is based on the center point. This algorithm requires three parameters, namely the number of *clusters*, *cluster* initialization, and system distance. The initial step taken in this method is to randomly select *k* objects as data centers in the data to be processed. The distance between the data center and the object is calculated by *Euclidean distance.* This algorithm *iteratively* increases the variation of values in each *cluster*. Then, the objects are placed in the closest group, and then calculated from the center point of the *cluster*. Next, it is determined if all data has been placed in the nearest *cluster*. This process is repeated until the midpoint of all *clusters is* formed and does not change anymore. Here is the *K-means Clustering* algorithm:

1. Step 1: Determine the number of *clusters k of* the dataset to be divided,
2. Step 2: randomly assign *k* data into *clusters*,
3. Step 3: Find the closest *cluster* of each data,
4. Step 4: Determine the extent center of each *cluster* and update the location of each *cluster* center to the new value of the extent center,
5. step 5: repeat step 3 to step 5 until the data in each *cluster is* centered.

2.3 Community Welfare

Community welfare is one of the benchmarks for the progress of a region. In a region, the level of community welfare can be assessed based on the human development index, household expenditure on food, household expenditure on education, household expenditure on health.

Household expenditure on food is very influential on community welfare. The higher the level of expenditure on food, the higher the welfare of the community, so that the community is considered to have been able to get out of the problem of hunger. If the community does not experience hunger, then the community will be more productive and can support the sustainability of life. Their income will increase and make the community prosperous.

The effect of spending on education is also very influential in the welfare of society. The higher the education, the higher the income. Another factor in welfare is health, which is closely related to education. A person who has a high level of education is not useful if he is not healthy, and a person is not necessarily prosperous without education. One of the important roles of welfare levels in the health factor is the ability to access health services. If there is less expenditure on health, then the degree of public health will increase according to the assessment of public welfare.

2.4 Regional Inequality

There are many types of inequality in a region. One of the most visible is inequality in economic development. According to Fauzi et al [3], the regional inequality index in Central Java decreased from 2006 to 2010..

Economic growth in a region greatly affects the progress of the region. The more the economy grows in the region, the more developed the region and its economy. So this case causes inequality between regions.

In addition to economic growth, agriculture is also very influential in regional inequality. The higher the concentration of a region on agriculture, the more visible the inequality. This is because regions that do not concentrate on agriculture will be left behind, making inequality more visible.

Factors underlying regional inequality include the quality of human resources, unemployment rate, level of health, health facilities, education facilities, infrastructure, investment, local revenue.

The level of community welfare is also seen from the number of people who work. The number of people who have jobs is a factor in the development of a region. The more people who have jobs, the higher the development of the region. So that it can reduce inequality between regions.

Based on the theory above, the following indicators will be used to measure regional welfare inequality:

1. Socio-economic: GRDP per capita, GRDP contribution of manufacturing industry, domestic investment, foreign investment, percent unemployment, percent poverty, number of criminal cases (offenses).
2. Health: percent of population with health problems, number of malnourished infants, ratio of hospital beds per 1000 population, percent of hospitals.
3. Education: student teacher ratio, average years of schooling, population ≤ 15 years of higher education graduates, school enrollment rate.

Results and Discussion

3.1 Data

The data used is 2020 data. The aspects that form the background of this data are geography, health, education, social and welfare. Table 1 below is the data that will be used in welfare grouping.

Table 1: Data on density, rate, poverty, poor families, diarrhea cases, number of clinics, number of hospitals, number of high schools and vocational schools in Surakarta sub-districts.

District	Density	The pace	Pre prosperous	Poor	Diarrhe a	Clinic	Hos- pital	HIGH SCHO OL	SMK
Laweyan	10245,83	0,27	1171	288	11498	25	5	6	6
Serengan	14977,43	0,88	739	273	1064	7	0	3	3
Pasar Kliwon	16289,83	0,54	1653	361	1210	6	3	3	3

Jebres	11031,4	0,05	3063,00	746,00	1913	17	3	5	5
Banjarsari	11395,68	0,68	3234	1075	2623	33	4	10	12

Data on population density and population growth rate represent the geographical success of a sub-district, while data on the number of pre-prosperous families and the number of first priority poor people represent social and welfare. One of the health problems that is used as a reference for poor health in Surakarta City is diarrhea disease, because the diarrhea morbidity rate in each sub-district is very high while in terms of facilities, clinics and hospitals are taken because they are the first reference for people in finding the nearest health facility. In terms of education, SMA and SMK facilities can be said to support 12 years of compulsory education.

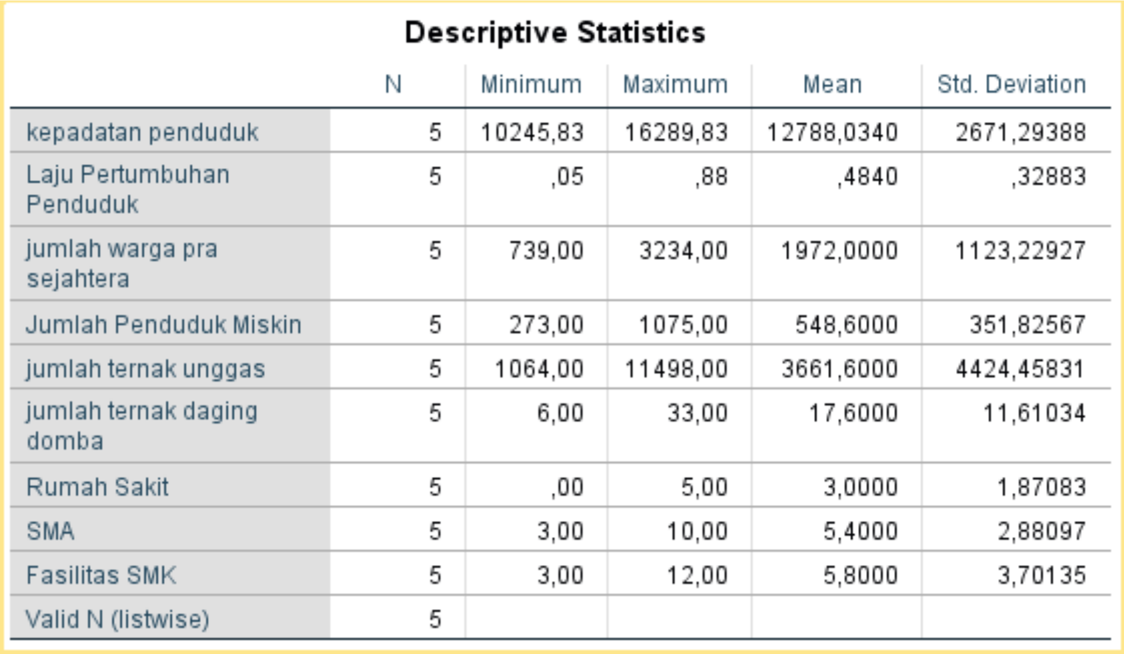
3.2 Discussion

The data collected has unit variables, so the data needs to be standardized. The standardization used is the transformation of the relevant variables into a Z value table.

Table 2: Standardized Data Table

ZDensity	ZPace	ZPre prosperous	ZPoor	Z Diarrhea	Z Clinic	ZHospital	ZHigh School	ZSMK
-,95168	-,65079	-,71312	-,74071	1,77115	,63736	1,06904	,20826	,05403
,81960	1,20427	-1,09773	-,78334	-,58710	-,91298	-1,60357	-,83305	-,75648
1,31090	,17030	-,28400	-,53322	-,55410	-,99911	,00000	-,83305	-,75648
-,65760	-1,31983	,97131	,56107	-,39521	-,05168	,00000	-,13884	-,21614
-52123	,59605	1,12355	1,49620	-,23474	1,32640	,53452	1,59668	1,67506

In addition to the transformation of the Z value, the data will be described, the results of the data description are in the table below. The expected *clusters* in this experiment are 2 *clusters*.

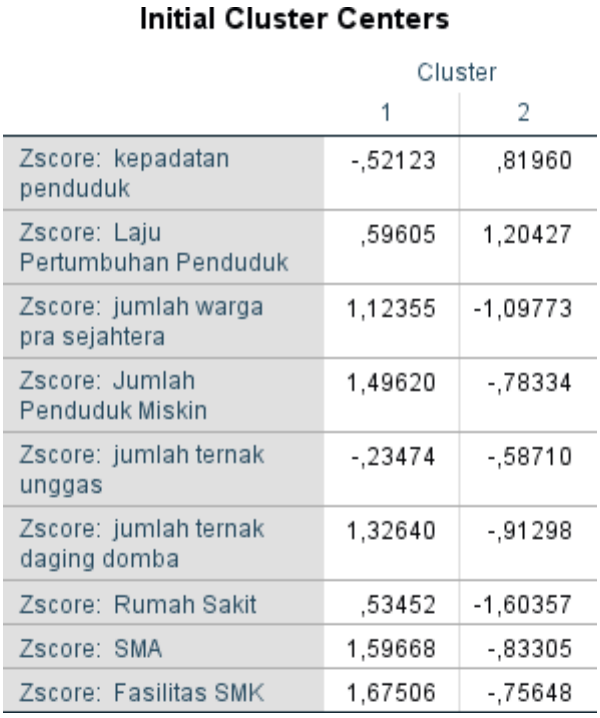


Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
kepadatan penduduk	5	10245,83	16289,83	12788,0340	2671,29388
Laju Pertumbuhan Penduduk	5	,05	,88	,4840	,32883
jumlah warga pra sejahtera	5	739,00	3234,00	1972,0000	1123,22927
Jumlah Penduduk Miskin	5	273,00	1075,00	548,6000	351,82567
jumlah ternak unggas	5	1064,00	11498,00	3661,6000	4424,45831
jumlah ternak daging domba	5	6,00	33,00	17,6000	11,61034
Rumah Sakit	5	,00	5,00	3,0000	1,87083
SMA	5	3,00	10,00	5,4000	2,88097
Fasilitas SMK	5	3,00	12,00	5,8000	3,70135
Valid N (listwise)	5				

Figure 1. Results of *descriptive statistics*

The description from Figure 1 will be used in the calculations in the *clustering* analysis.

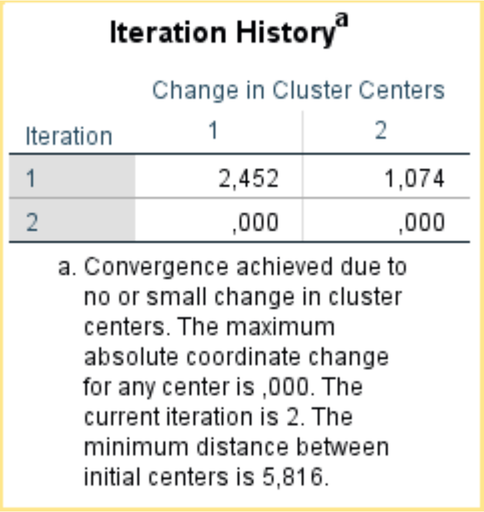


Initial Cluster Centers

	Cluster	
	1	2
Zscore: kepadatan penduduk	-,52123	,81960
Zscore: Laju Pertumbuhan Penduduk	,59605	1,20427
Zscore: jumlah warga pra sejahtera	1,12355	-1,09773
Zscore: Jumlah Penduduk Miskin	1,49620	-,78334
Zscore: jumlah ternak unggas	-,23474	-,58710
Zscore: jumlah ternak daging domba	1,32640	-,91298
Zscore: Rumah Sakit	,53452	-1,60357
Zscore: SMA	1,59668	-,83305
Zscore: Fasilitas SMK	1,67506	-,75648

Figure 2. The result of the first process in *Clustering*

Figure 2 illustrates the results of the first process or initial values in *clustering*. The data will then be processed into several iterations as the next step in *clustering*.



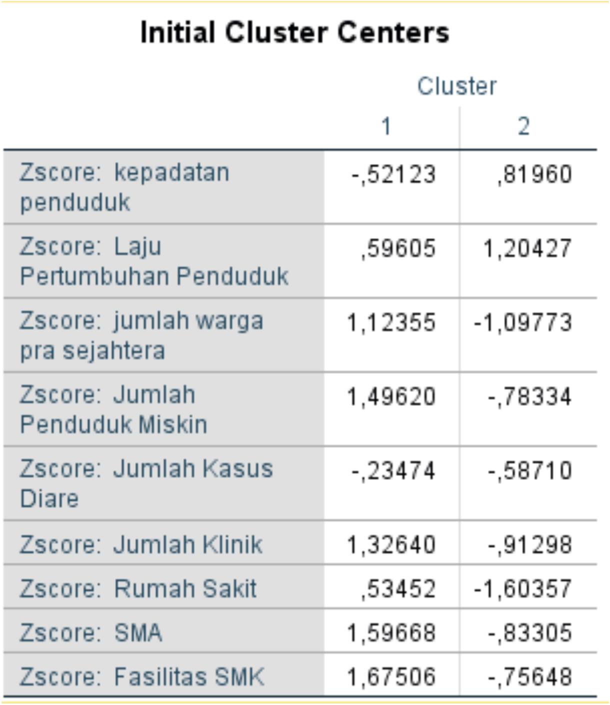
Iteration History[a]

Iteration	Change in Cluster Centers 1	2
1	2,452	1,074
2	,000	,000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is ,000. The current iteration is 2. The minimum distance between initial centers is 5,816.

Figure 3. Iteration result

Figure 3 explains the number of iterations that occur, many iterations that occur in this experiment to get 2 *clusters* there are 2 iterations, and the minimum distance between *cluster* centers formed from the iteration results is 5.816.



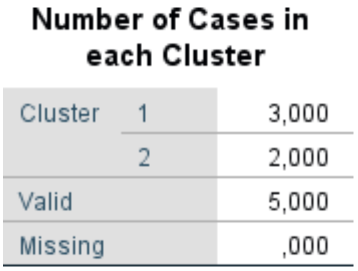
Initial Cluster Centers

	Cluster 1	2
Zscore: kepadatan penduduk	-,52123	,81960
Zscore: Laju Pertumbuhan Penduduk	,59605	1,20427
Zscore: jumlah warga pra sejahtera	1,12355	-1,09773
Zscore: Jumlah Penduduk Miskin	1,49620	-,78334
Zscore: Jumlah Kasus Diare	-,23474	-,58710
Zscore: Jumlah Klinik	1,32640	-,91298
Zscore: Rumah Sakit	,53452	-1,60357
Zscore: SMA	1,59668	-,83305
Zscore: Fasilitas SMK	1,67506	-,75648

Figure 4. Final *cluster* results

Figure 4 explains that of the 2 expected *clusters*. If the value that appears is negative, then in that *cluster*, the average of the existing data is below the average value, while if it is positive, it means it is above the average.

The conclusion obtained from the table above is that in the first *cluster*, the sub-district has a population density, and the number of diarrhea cases is below average, whereas the population growth rate in the number of poor people, the number of poor people, the number of diarrhea cases, the number of clinics, the number of hospitals, the number of high schools, and the number of vocational schools are above average.



Number of Cases in each Cluster

Cluster	1	3,000
	2	2,000
Valid		5,000
Missing		,000

Figure 5. Results of the number of sub-districts in each *cluster*

The number of sub-districts in each *cluster* is shown in Figure 5, *cluster* 1 consist of 3 sub-districts and *cluster* 2 consists of 2 sub-districts.

The results of this *clustering* show that those included in *cluster* 1 are Laweyan, Jebres, and Banjarsari sub-districts. While those included in *cluster* 2 are Serengan, and Pasar Kliwon.

2 Conclusion

Based on the results obtained from the application of *K-means Clustering* in grouping the welfare of citizens based on sub-districts in the city of Surakarta, the following conclusions can be drawn:

1. *cluster* 1 consists of 3 sub-districts Laweyan, Jebres, Banjarsari, where these three sub-districts have a population density level and the number of diarrhea cases that are below average, so these three sub-districts have a good level of health and a good level of population density, but in the number of poor people, the number of poor people is still lacking, even though health and education facilities are adequate, and
2. *cluster* 2 consists of 2 sub-districts, Pasar Kliwon and Serengan. these two sub-districts have a below average number of poor and pre-prosperous people so it can be said that these sub-districts have good welfare, but with good welfare, health and education facilities are still less than the other 3 sub-districts.

3 Advice

Based on the results obtained from the application of *K-Means Clustering* that we did in grouping the welfare of citizens based on sub-districts in Surakarta City, there are several suggestions that can be used for future improvements, namely:

1. this method can be applied to multiple regions including sub-districts, districts, cities, and provinces, and
2. This method can serve as an evaluation for local governments to combat social inequality.

Reference List

[1] Badan Pusat Statistik. 2021. *Kota Surakarta Dalam Angka 2021*. Magelang : Badan Pusat Statistik.

[2] Hidayat, R., Wasono, R., dan Darsyah, M.Y. . 2017. *Pengelompokkan Kabupaten/Kota di Jawa Tengah Menggunakan Metode K-means dan Fuzzy C-Means*. Seminar Nasional Pendidikan Sains dan Teknologi Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Muhammadiyah Semarang. 240-250.

[3] Fauzi, M. R., Rustiadi, E., dan Mulatsih, S. .2019. *Ketimpangan Pola Spasial dan Kinerja Pembangunan Wilayah di Provinsi Jawa Timur.* Journal of Regional and Rural Development Planning. 161.